Journal of Applied Informatics and Computing (JAIC)
Vol.10, No.2, April 2026, pp. 1530~1540
e-ISSN: 2548-6861

Principal Component Analysis (PCA) for Interval-Valued Symbolic Data: A Comparison of the Center and Vertex (TOPS) Methods

Boono Yaba Benjamin [1]*, Mabela Makengo Rostin [2]*, Kikomba Kahungu Michael [3], Mbuyi Lunkondo Patience [4]***, Kipulu Ngimbi Serge [5]******

* Department of Mathematics, Statistics and Computer Science, University of Kinshasa
** Department of Exact Sciences, Higher Pedagogical Institute of Gombe–Kinshasa
*** Center for Interdisciplinary Research, National Pedagogical University (CRIDUPN)
**** Computer Science Section, Higher Pedagogical and Technical Institute of Kinshasa
benjamin.boono@unikin.ac.cd [1], rostin.mabela@unikin.ac.cd [2], michaelkikomba01@gmail.com [3], patiencembuyi95@gmail.com [4], sergekipulu@gmail.com [5]

Article Info

Article history:

Received 2026-01-29
Revised 2026-02-21
Accepted 2026-04-10

Keywords:

Symbolic Data Analysis (SDA),
Principal Component Analysis (PCA),
Centers,
Vertices (TOPS),
Dimensionality Reduction.

ABSTRACT

Classical dimensionality reduction techniques, such as Principal Component Analysis (PCA), are widely used to explore the structure of multivariate datasets. However, these methods are traditionally restricted to situations in which each variable is represented by a single numerical value per individual. The emergence of symbolic data, particularly interval-valued data, has introduced new challenges in the field of data science. In this framework, a single variable may take multiple possible values, reflecting either measurement uncertainty or intrinsic variability of the observation. Such data therefore provide a more faithful representation of the complexity of observed phenomena, but they require specifically adapted analytical methodologies. This paper aims to compare two PCA variants applied to interval-valued symbolic data: The Center Method, in which each interval is represented by its midpoint, and the Vertex Method (TOPS), in which the lower and upper bounds of each interval are jointly exploited. We formally define interval-valued variables, present the algorithmic steps of both the Center and TOPS methods, analyze their computational complexity, and introduce evaluation metrics including explained variance, reconstruction error, and sensitivity analysis with respect to interval width. The objective is to assess the extent to which these approaches preserve the information contained within intervals and to determine which method proves more appropriate for a given dataset. Using a biomedical dataset (n = 1021 individuals, p = 7 interval-valued variables), we show that while the Center method provides strong dimensional condensation and interpretability, the TOPS method more faithfully preserves the geometry of intervals in the presence of high variability. This study clarifies the theoretical differences between the two approaches and proposes a systematic evaluation framework for interval-valued symbolic PCA methods.

I. INTRODUCTION

In a scientific context increasingly characterized by the explosion of data volumes and the growing complexity of data acquisition systems, new forms of data representation continuously emerge. Among these, symbolic data constitute a particularly important framework. Symbolic Data Analysis (SDA) extends classical statistical methodology to complex data types, including variables represented in interval form.

In many biomedical, environmental, social, security, and technical contexts, measurements are naturally represented as intervals due to:

- Measurement uncertainty,
- Data aggregation,

- Confidentiality constraints,
- Or repeated measurements.

Classical PCA assumes point observations (a single scalar value per individual), which necessitates methodological adaptation when dealing with interval-valued data. Symbolic data represent objects or individuals described by multiple values, internal structures, or complex descriptions [9]. Formally, a symbolic variable X is defined such that for each statistical unit iii, it takes non-atomic values. In contrast to classical datasets, where each individual is described by a single numerical value per variable, symbolic data allow richer representations. These types of data are increasingly important in data science, particularly when describing imprecise, aggregated, or uncertain objects [10].

Symbolic data may take several forms, including:

- An interval of values (e.g. height between 150-190 cm),
- A set of categories (e.g. preferred colors = {blue, green, red}),
- A probability distribution (e.g. 30% red, 70% yellow),
- A textual or logical description.

It is now common in health sciences, environmental studies, and engineering to encounter variables characterized not by a single scalar value, but by an interval of the form $[a_i^-, a_i^+]$, which reflects either measurement uncertainty or deliberate aggregation of repeated observations [10].

Four major families of symbolic data can be distinguished:

1. Interval-valued variables A variable takes an interval $[a_i^-, a_i^+]$ of continuous values. Example: body temperature between 36.5°C and 38°C.
2. Set-valued (multi-valued) variables A variable may take a set of modalities. Example: a patient presenting symptoms {fever, cough, nausea}.
3. Histogram or distribution-valued variables Each variable is represented by a frequency distribution over its modalities. Example: distribution of children's sleep time: 50% night, 30% nap, 20% awake.
4. Modal (weighted categorical) variables Each category is associated with a weight or probability. Example: a celebrity's hair color = {red (0.7), blue (0.3)}.

Symbolic data are particularly useful when:

- Data originate from grouped observations (e.g., summaries of patient groups or political affiliations),
- Measurement uncertainty or imprecision is present,
- One seeks to model reality more faithfully by accounting for internal variability within each statistical unit.

They are particularly relevant when information is extracted from textual data, sensor systems, or expert systems [11].

TABLE 1.
INTERVAL-VALUED, SET-VALUED, AND MODAL SYMBOLIC DATA

Patient	Temperature (°C)	Possible symptoms	Pain distribution
K	[36.5, 40.5]	{fever, fatigue}	{mild (0.4), severe (1.0)}
M	[38.0, 39.2]	{fever, cough, nausea}	{moderate (0.6)}

These new types of data pose significant challenges to classical statistics—both descriptive and inferential—which fundamentally rely on assumptions of precision and linearity. In particular, traditional dimensionality reduction techniques (factorial methods), such as Principal Component Analysis (PCA), must be rethought in order to accommodate the symbolic nature of interval-valued data.

Throughout this work, we aim to address the following fundamental questions:

- How can an interval be projected onto a factorial space?
- Which information should be preserved: the midpoint, the bounds, or the entire internal structure?

To answer these questions, several extensions of PCA have been proposed in the literature, notably those developed by [2], [3], and [7]. These approaches attempt to adapt the geometric and algebraic foundations of PCA to the constraints imposed by interval-valued symbolic data while preserving as much information as possible regarding their internal variability.

The present research builds upon existing work in Symbolic Data Analysis (SDA) applied to PCA for interval-valued data. It provides several methodological and empirical contributions that strengthen the comparative understanding of existing approaches.

1. Formal Mathematical Comparison between Center-Based Symbolic PCA and Vertex-Based (TOPS) Symbolic PCA

Unlike studies that present these two methods independently or descriptively, our work proposes a rigorous and parallel formalization of both approaches.

We explicitly establish:

- The transformation of interval-valued data into a point representation (Center Method) via

$$c_{ij} = \frac{a_{ij} + b_{ij}}{2}$$

the structural transformation through concatenation of the lower and upper bounds (TOPS method), $X^* = [L|U]$

This formalization enables:

1. A coherent comparison of the resulting covariance matrices,
2. A geometric analysis of the produced factorial spaces,

3. A clear understanding of inter-object and intra-object dispersion differences.

Thus, the contribution does not merely consist of an empirical comparison, but rather relies on a structured theoretical framework that highlights the geometric and statistical implications of each transformation.

2. Explicit Algorithmic Presentation to Ensure Reproducibility

Within an open science and computational reproducibility perspective, we provide a detailed algorithmic description of both methods in the form of systematic procedural steps:

a. Computation of the transformations (centers),
b. Standardization,
c. Construction of covariance matrices,
d. Spectral decomposition,
e. Factorial projection.

This algorithmic structuring allows:

a. Direct implementation in Python,
b. Clear estimation of computational costs,
c. Exact reproducibility of experimental results.

It addresses a frequent gap in the literature, where symbolic extensions of PCA are often presented conceptually without complete procedural specification.

3. Computational Complexity Analysis

An important contribution of this study lies in the comparative evaluation of algorithmic costs.

We demonstrate that:

- Center-based PCA preserves the original dimensionality p, avec une complexité asymptotique : $o(np^2)$
- whereas the vertex-based PCA (TOPS) doubles the dimensionality from p to $(2p)$, which results in a computational complexity of $O(n(2p)^2)$ that is, $4O(np^2)$.

This analysis highlights a fundamental trade-off:

- The Center method prioritizes computational efficiency,
- The TOPS method prioritizes geometric richness at the cost of increased computational burden.

By integrating this computational dimension, our study provides an additional decision criterion for methodological selection, depending on dataset size and structural complexity.

4. Introduction of Reconstruction Error as an Evaluation Metric

Beyond the traditionally used explained variance in PCA, we introduce the reconstruction error as a quantitative measure of fidelity. $RE = \| X - \hat{X} \|^2_F$

This metric allows:

- Evaluation of each method's ability to reconstruct the original interval structure,
- Objective comparison of the information loss induced by the transformation process,
- Identification of situations in which center-based simplification leads to significant degradation.

The introduction of this measure enriches the evaluation framework beyond purely geometric indicators and provides a more robust empirical validation dimension.

5. Sensitivity Analysis with Respect to Interval Width

Finally, we propose a sensitivity analysis based on the artificial variation of interval radi:

$$r^{\alpha}_{ij} = \alpha r_{ij} \;\; \alpha > 0$$

This approach makes it possible to investigate:

- The stability of eigenvalues,
- The robustness of factorial axes,
- The variation in explained variance,
- The evolution of reconstruction error.

This analysis constitutes an original contribution of the present work, as few studies assess the robustness of symbolic PCA methods under amplified uncertainty conditions.

The objective of this article is to compare two main PCA approaches for purely interval-valued symbolic data:

1. The Center-based PCA method, which represents each interval by its midpoint while implicitly adjusting variances according to internal dispersion; and
2. The Vertex-based PCA method (TOPS), which represents each object by the extreme points defined by all combinations of the lower and upper bounds of its intervals.

The remainder of this article is organized as follows:

- Section 2 presents interval-valued symbolic data and their fundamental properties;
- Section 3 details the PCA methods adapted to interval-valued symbolic data;
- Section 4 describes the experimental methodology implemented for both approaches;
- Section 5 presents the obtained results and their interpretation;
- Section 6 discusses limitations and future perspectives; and finally,
- Section 7 concludes the study by summarizing the methodological contributions and comparative findings.

II. INTERVAL-VALUED SYMBOLIC DATA

A. Formal Definition

Interval-valued symbolic data constitute a generalization of classical data. Instead of associating each statistical unit with a real value x_i, a symbolic variable X_i is defined as a closed interval:

$$X_i = [x_i^-, x_i^+] \; with \; x_i^- \leq x_i^+$$

This form of representation naturally applies to several contexts. For example, in hospital databases, a patient's age may be recorded as a range (eg. [20,30]), resulting from data aggregation or deliberate anonymization. In meteorology, daily temperature may be represented by its extremes: minimum temperature T_{min} and maximum temperature T_{max}. Formally, an interval-valued symbolic variable is defined as follows:

$$X: i \rightarrow [x_i^-, x_i^+]$$

It is defined for each observation $i \in \{1, \ldots, n\}$, and is entirely characterized by its lower and upper bounds.
Two fundamental characteristics can be derived from this representation:

$$center\ (midpoint): c_i = \frac{x_i^- + x_i^+}{2},$$

$$radius\ (half - range):\ r_i = \frac{x_i^+ - x_i^-}{2}$$

Let $X = (X_{ij})$ be an interval-valued data matrix with n observations and p interval-valued variables, where each entry is $X_{ij} = [a^-{}_{ij}, a^+{}_{ij}], i = 1, \ldots, n$;

$$j = 1, \ldots, p \; with \; a^-{}_{ij} \leq a^+{}_{ij}$$

Define the center (midpoint) matrix $C = (c_{ij}) \in \mathbb{R}^{n \times p}$ and the radius (half-range) matrix $R = (r_{ij}) \in \mathbb{R}^{n \times p}$ by $c_{ij} = \frac{a_{ij}^- + a_{ij}^+}{2}, r_{ij} = \frac{a_{ij}^+ - a_{ij}^-}{2}$ Then each interval can be written equivalently as $X_{ij} = [c_{ij} - r_{ij}, c_{ij} + r_{ij}]$, so that the whole interval-valued matrix admits the compact decomposition $X = C \pm R$. More explicitly, the lower- and upper-bound matrices are recovered as $A^- = C - R$, $A^+ = C + R$, where $A^- = (a_{ij}^-) \; and \; A^+ = (a_{ij}^+)$ Equivalently, for all (i, j), $a_{ij}{}^- = c_{ij} - r_{ij}$, $a_{ij}{}^+ = c_{ij} + r_{ij}$.

These components can be exploited within different analytical frameworks, as will be discussed in the subsequent sections.

B. Specific Properties

Unlike classical data, interval-valued symbolic data exhibit specific statistical and geometrical characteristics:

- Internal dispersion: each observation contains its own uncertainty or intrinsic variability.
- Lack of unique representability: there is no single point that can perfectly summarize an interval, as is the case for a scalar value.
- Potential asymmetry: although most approaches rely on symmetric centers, some intervals may reflect biases (for instance, the interval [5,10] is not equivalent to [10,15] in terms of their impact on certain measures).

These properties call for a reconsideration of classical statistical tools, particularly those that exploit the geometric structure of data.
It is within this framework that Principal Component Analysis (PCA) adapted to interval-valued symbolic data becomes relevant.

III. SYMBOLIC PCA METHODS FOR INTERVAL-VALUED DATA

Principal Component Analysis (PCA) is a classical dimensionality reduction technique based on the computation of directions that maximize the variance of the data. When applied to interval-valued symbolic data, several adaptations are possible. In this paper, we present two major approaches that are widely described in the literature. [2], [5], [7]

A. Center-Based PCA

This approach, proposed in the works of Billard and Diday [3], also relies on the centers cic_ici, while retaining the radii rir_iri as indicators of dispersion (radius-based adjustment) :

$$r_i = \frac{x_i^+ - x_i^-}{2}$$

This information can thus be incorporated into the variance–covariance matrix, for instance by adjusting the weights or by considering the half-widths as complementary variables.
The input matrix is therefore enriched by the internal structure of the intervals $\tilde{X} = \{c_i, r_i\}$.

The center-based PCA method introduces variability into the analysis without excessive algorithmic complexity. It allows a more faithful projection of the actual dispersion of symbolic objects. However, the joint interpretation of centers and radii requires particular care and may be sensitive to the scale of the variables, which can sometimes make the analysis challenging.

B. Vertex-Based PCA (TOPS)

The TOPS method (Two Opposite Points Strategy) is one of the most refined approaches proposed for interval-valued symbolic PCA. It consists in representing each variable by its two extreme bounds and each object by the 2^p combinations of these bounds. This construction generates a cloud of points associated with each symbolic object. i :

$$\mathcal{N}_i = \left\{ (x_{i1}^{*}, \quad x_{i2}^{*}, \dots, x_{ip}^{*}) | x_{ij}^{*} \in \{x_{ij}^{-}, x_{ij}^{+}\} \right\}$$

Each object thus becomes a cloud of points in the factorial space $\mathbb{R}^p$, whose principal component analysis provides a comprehensive representation of its geometric structure. This method faithfully restores the shape and dispersion of each object and is particularly well suited when uncertainty in the values is significant or heterogeneous. It should be noted that this approach involves an exponential computational complexity (2^p vertices per object), which may become computationally expensive for high dimensional datasets.

The method also requires the introduction of specific distance measures (e.g. Hausdorff and Wasserstein distances) to enable a finer comparison between objects [4], [9], [17].

In the following, we compare these two methods through an experimental study conducted on real data comprising 1021 individuals described by seven (7) interval-valued symbolic variables, by analyzing their behavior according to several quality criteria.

IV. MÉTHODOLOGY

In order to rigorously compare the three Principal Component Analysis approaches adapted to interval-valued symbolic data namely the classical PCA, the center based PCA, and the vertex based (TOPS) PCA we designed a methodology that allows precise control of the internal characteristics of the intervals and enables the assessment of the robustness of these three symbolic PCA methods under realistic conditions.

A. Data Description

For the application, we used a dataset collected in a biomedical context, concerning the morphological and clinical characteristics of patients followed between 2018 and 2021. The data were collected within the Gynecology Department of the Sino-Congolese Friendship Hospital (HASC), located in Kinshasa, in the Tshangu district, N'djili municipality, as well as at the Kintambo Maternity Hospital, located in the Kintambo municipality of Kinshasa.

B. Comparison Criteria

To assess the relevance of each PCA method applied to interval-valued symbolic data, we retained four main criteria:

1. Variance explained by the leading components. A good dimensionality reduction method should condense as much information as possible into the first principal components (in particular, the first two). We therefore compare the cumulative relative inertia explained by the first two axes.
2. Projection fidelity. For each method, we measure the distance between the projected symbolic object and its original representation, using classical metrics, namely the L2, Hausdorff, and Wasserstein distances.
3. Interpretability of the components. PCA should provide a clear and meaningful interpretation of the factorial axes. We thus evaluate the stability of variable contributions, the clarity of correlation circles, and the readability of the resulting groupings.
4. Algorithmic complexity. The computational time, memory consumption, and scalability of each method are empirically assessed.

Each of these criteria is discussed in the following section based on the results obtained from the application.

V. RESULTS AND INTERPRETATION

The two Principal Component Analysis methods adapted to interval-valued symbolic data were applied to our dataset (Section A). The analyses were conducted according to the methodological criteria defined in Section 4.2, using Python 3 with Jupyter Notebook 7.0.8.

A. Variance Explained By The Leading Components

TABLE 2.
SUMMARY OF INERTIAS FOR THE TWO METHODS

Axis	Center-Based Inertia (%)	Vertex-Based (TOPS) Inertia (%)
PC1	57.90	26.17
PC2	20.55	19.05
PC3	12.51	14.09
PC4	7.76	12.49
PC5	1.05	11.53
PC6	0.22	9.19
PC7	0.01	7.48

The center-based PCA produces a very high inertia on the first principal axis, explaining approximately 57.9% of the total variance. In contrast, the vertex-based (TOPS) PCA exhibits a much more distributed inertia structure, with the first axis accounting for only 26.17% of the total variance.

Considering the first two principal components, the center-based method concentrates more than 75% of the total information, indicating a strong capacity for dimensionality reduction. This behavior reflects the fact that representing each interval by its center captures the dominant trends of the data while smoothing internal variability.

On the other hand, the vertex-based method spreads the information over a larger number of components, as the variability induced by the interval bounds is explicitly preserved. This more gradual decay of inertia highlights the greater dispersion associated with the extreme values of the intervals and suggests that higher-dimensional representations are required to adequately summarize the data structure.

Overall, these results show that the center-based PCA is more efficient for dimensionality reduction, whereas the vertex-based PCA provides a richer but more dispersed representation, better capturing the internal variability of interval-valued symbolic data.

B. Projection Fidelity (Average Quality Of Representation)

TABLE 3.
PROJECTION FIDELITY ON THE PRINCIPAL PLANE

Method	Average quality (cos^2)
Center-based	0.6537
Vertex-based (TOPS)	0.3652

The projection of symbolic objects onto the principal factorial plane is more faithful for the center-based PCA, as indicated by a relatively high average cos^2 value (0.6537). This result suggests that the first two principal components provide a good approximation of the original data structure when intervals are represented by their centers.

In contrast, the vertex-based (TOPS) method exhibits a lower projection fidelity ($cos^2 = 0.3652$). This behavior is a direct consequence of the explicit consideration of the internal variability of intervals through their extreme bounds. By preserving more detailed geometric information, the TOPS method distributes the variance across a larger number of components, which reduces the quality of low-dimensional projections.

These results highlight a classical trade-off in symbolic PCA: while the center-based approach favors compact and accurate low-dimensional representations, the vertex-based method prioritizes the preservation of internal variability at the expense of projection fidelity on the first factorial plane.

C. Interpretability Of The Principal Components

TABLE 4.
READABILITY OF FACTORIAL AXES AND CORRELATION CIRCLES

Method	Clarity of correlation circles and groupings
Center-based	Very good
Vertex-based (TOPS)	Moderate

The center-based PCA yields highly readable factorial axes, with well-defined correlation circles and clearly interpretable groupings of variables. This clarity stems from the reduced complexity of the representation, where each symbolic object is summarized by its center, leading to more stable and interpretable loadings.
Conversely, the vertex-based PCA (TOPS) produces correlation structures that are less clear, due to the increased dispersion induced by the multiple vertices associated with each object. Although this method provides a richer geometric representation, it may complicate the interpretation of the principal components, especially in low-dimensional visualizations.

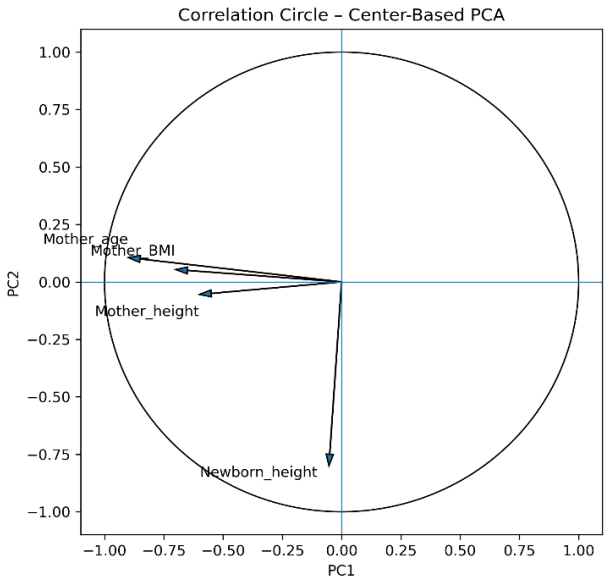

Figure 1. Correlation circle center-based PCA.

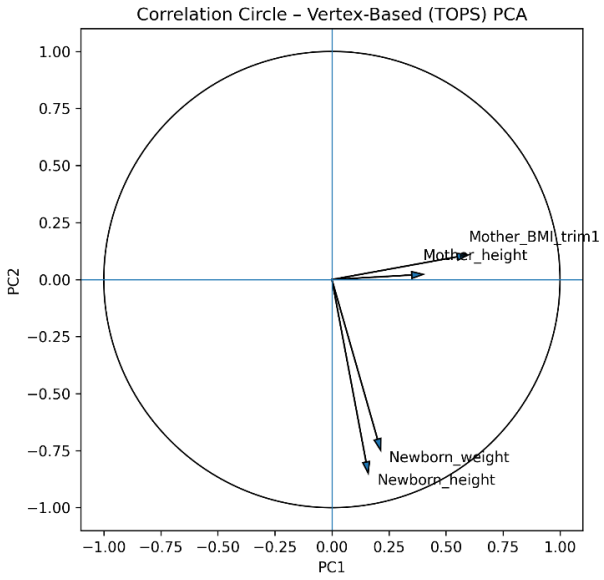

Figure 2. Correlation Circle Vertex-Based (TOPS) Method

In Principal Component Analysis, the correlation circle facilitates the visualization of the contributions of variables to the principal components. The weights observed on the first two axes for each PCA method represent the amount of information to be interpreted along each axis. The center-based method provides a clear and well-structured visualization of the correlation circle (Figure 1). The arrows are well oriented, and the groupings of variables are clearly identifiable.

In contrast, the **v**ertex-based (TOPS) method exhibits a more scattered correlation circle, with less readable axes (Figure 2). This dispersion reflects the greater variability induced by the use of interval bounds, which complicates the geometric interpretation of the factorial structure.

TABLE 4.
VARIABLE CONTRIBUTIONS ON THE FIRST TWO AXES CENTER-BASED METHOD

Variable	PC1 (Center)	PC2 (Center)
Mother's age (years)	0.227937	0.001961
Mother's weight (1st trimester, kg)	0.222456	0.002731
Mother's weight (3rd trimester, kg)	0.243891	0.001153
Mother's BMI	0.240992	0.001097
Mother's height (m)	0.057890	0.000375
Newborn's weight (kg)	0.006046	0.485510
Newborn's height (cm)	0.000787	0.507174

The first principal component (PC1) is mainly driven by maternal characteristics, including mother's age, maternal weight (first and third trimesters), and BMI. These variables exhibit very high contributions, indicating strong correlations and a dominant common factor associated with maternal morphology. The second principal component (PC2) is primarily explained by newborn characteristics, namely birth weight and newborn height, which show high positive contributions on this axis. Overall, the correlation circle obtained with the center-based PCA is clear and well structured. Variables are strongly correlated within their respective groups, and the factorial axes exhibit a high degree of stability, allowing for a straightforward and reliable interpretation.

TABLE 5.
VARIABLE CONTRIBUTIONS ON THE FIRST TWO AXES VERTEX-BASED (TOPS) METHOD

Variable	PC1 (TOPS)	PC2 (TOPS)
Mother's age (years)	0.347797	0.010128
Mother's weight (1st trimester, kg)	0.266835	0.013157
Mother's weight (3rd trimester, kg)	0.148337	0.007536
Mother's BMI	0.176977	0.008078
Mother's height (m)	0.018306	8.2373E-05
Newborn's weight (kg)	0.032129	0.461402
Newborn's height (cm)	0.009618	0.499616

As with the center-based approach, the same variable groupings are observed with the vertex-based (TOPS) method, distinguishing maternal characteristics from newborn characteristics. However, the magnitude of the coefficients is systematically lower.

In the correlation circle, the arrows are shorter, indicating a more diffuse representation of the variables on the first two axes. This behavior reflects the greater internal variability captured by the TOPS method through the use of interval bounds. While this approach preserves richer geometric information, it results in moderate axis stability and makes the interpretation of the factorial structure more challenging.

C. Algorithmic Complexity

The study of algorithmic complexity makes it possible to compare the computational cost of the two symbolic PCA approaches (center-based and vertex-based methods). Let n denote the number of individuals and p the number of symbolic variables.

1. Center-Based Method

Each interval is represented by its center $\frac{min+max}{2}$. The size of the data matrix remains $n \times p$. The computation of the centers has complexity $\mathcal{O}(np)$, which is negligible compared to the dominant costs $\mathcal{O}(np^2 + p^3)$.

2. Vertex-Based (TOPS) Method

In the vertex-based (TOPS) PCA method, each interval-valued variable is represented by its two extreme bounds. Consequently, each symbolic object described by p interval-valued variables is expanded into 2^p vertices corresponding to all possible combinations of lower and upper bounds. The resulting expanded data matrix therefore has size $(n2^p)p$. The computational cost of this method is dominated by:

- the generation of the vertices, with complexity $\mathcal{O}(n2^p p)$, and
- the construction of the variance–covariance matrix and its eigen-decomposition, whose cost becomes

$$\mathcal{O}(n2^p p^2 + p^3)$$

As a result, the overall algorithmic complexity grows exponentially with the number of variables p. This exponential growth leads to a significant increase in both computational time and memory usage, making the TOPS method computationally expensive for high-dimensional interval-valued datasets.

Despite this cost, the TOPS approach provides a more faithful geometric representation of symbolic objects by fully preserving their internal variability through the use of extreme points.

Due to the generation of extreme points, the vertex-based (TOPS) method significantly increases both memory requirements and computational time. While the vertex-based approach provides a better representation of the internal dispersion of symbolic objects, this advantage comes at the cost of a reduced concentration of information on the leading principal components.

The computational complexity of the TOPS method grows exponentially with the number of variables, which makes it difficult to apply to large-scale datasets.

Nevertheless, it remains manageable for moderate-dimensional settings (typically $p < 10$).

D. Projection Fidelity

The resulting projections reveal significant differences between the two approaches. The center-based method projects each observation onto a single point, without explicitly accounting for internal dispersion. In contrast, the vertex-based (TOPS) method generates clouds of points around each observation, representing all possible combinations of the interval bounds.

E. Interpretability Of The Principal Components

The correlation circles obtained for the two methods also reveal significant differences. For the center-based method, the vectors are strongly aligned with the first principal axis, indicating a highly correlated structure among the variables.

In contrast, the vertex-based (TOPS) method displays more dispersed vectors, illustrating a more balanced representation of inter-variable variance across the factorial axes. With regard to K-means clustering, both methods identify similar underlying structures. However, the clusters appear more diffuse when using the vertex-based (TOPS) method, as illustrated in Figures 3 and 4. This behavior reflects the greater internal variability preserved by the TOPS representation.

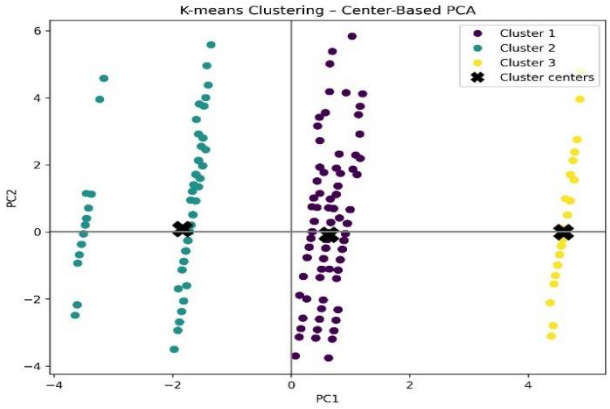

Figure 3. K-means clustering plot Center-Based PCA

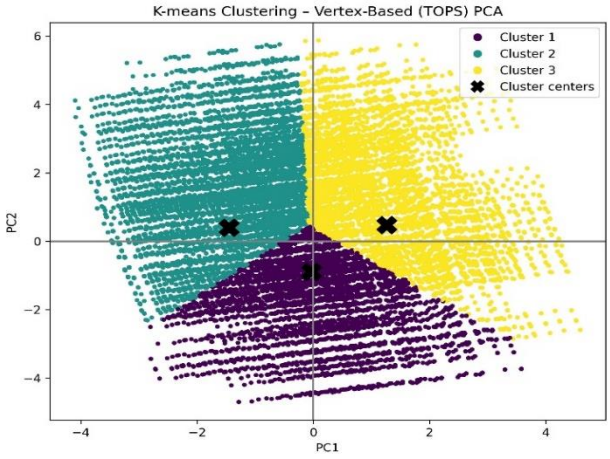

Figure 4. Graphique K-Means de la méthode de sommets

VI. Discussion Limitations And Future Perspectives

The results reveal substantial structural differences between the Center-based PCA and the Vertex-based PCA (TOPS), in terms of both geometric properties and computational behavior.

A. Interpretability And Dimensional Condensation: An Advantage Of The Center-Based Method

Center-based PCA exhibits a strong capacity for dimensional condensation. In our application, the first principal component explains nearly 58% of the total inertia, while the first two components cumulatively account for more than 75% of the total information. This result reflects a strong structural correlation among the variables once transformed by their midpoints.

From a geometric standpoint, the center representation maps each interval to a single point in $\mathbb{R}^p$. This transformation leads to:

- A concentration of variance on a limited number of principal axes,
- A clear structuring of the correlation circles,
- Stable variable contributions across components.

The observed contributions indicate that the first principal component (PC1) is predominantly driven by maternal variables (age, weight, BMI), whereas the second component (PC2) is mainly associated with newborn characteristics.

This clear separation greatly facilitates substantive interpretation of the factorial axes.

The average quality of representation (mean $\cos^2$= 65%) further confirms that the projections obtained are faithful within the reduced factorial space. Therefore, the Center method proves particularly effective when the primary objective is:

- Data visualization,
- Information synthesis,
- Classification or clustering (as illustrated by the more structured K-Means partitioning).

However, this performance is partly explained by a structural simplification: each interval is summarized by its midpoint, thereby explicitly neglecting the internal variability captured by the radii. In other words, intra-object dispersion is partially absorbed into the mean representation, which may lead to information loss when intervals are wide or heterogeneous.

B. Preservation Of Internal Variability: A Key Strength Of The Tops Approach

The TOPS method adopts a geometrically richer approach by representing each variable through its lower and upper bounds, and each object through the set of all possible extreme combinations.

This construction allows:

- Explicit preservation of the internal structure of intervals,
- Integration of intra-object dispersion into the analysis,
- Recovery of the complete geometric configuration of each observation.

Empirically, this results in a more dispersed distribution of inertia. The first principal component explains only approximately 26% of the total variance, and the information is distributed across a larger number of principal components.

This phenomenon should not be interpreted as a methodological weakness; rather, it reflects the richness of the preserved information. Indeed, the internal variability of the intervals generates a more complex structure in the factorial space, thereby preventing excessive condensation onto a single axis.

However, this geometric richness entails several implications:

a. The correlation circles appear more diffuse.
b. The vectors are shorter, reflecting increased dispersion.
c. Individual variable contributions are smaller, making interpretation less immediate.
d. The average projection quality (mean $\cos^2$=36%) is lower, since each object is no longer represented by a single point but by an implicit cloud of extreme configurations.

Thus, the TOPS method prioritizes structural fidelity at the expense of interpretative simplicity.

C. Geometric-Statistical Trade-Off In Interval-Valued PCA

Both methods embody a fundamental trade-off in symbolic data analysis:

a. The Center method maximizes dimensional condensation and interpretability.
b. The TOPS method maximizes the preservation of
c. uncertainty and internal variability.

When intervals are relatively narrow and homogeneous, the Center method provides a sufficiently accurate approximation with substantial computational gains. Conversely, when:

a. The radius vary considerably across individuals,
b. Uncertainty is structural rather than incidental,
c. Internal dispersion constitutes a central analytical component, the TOPS method becomes more appropriate, as it avoids premature reduction of the intrinsic complexity of the objects.

D. Computational Implications

From an algorithmic perspective, the TOPS method doubles the data dimensionality ($from\ p$ à $2p$), which results in an asymptotic complexity approximately four times higher than that of the Center method:

$$O(n(2p)^2) = 4O(np^2)$$

In our study ($p = 7$), this computational burden remains manageable. However, for higher-dimensional datasets, the TOPS method may become demanding in terms of both memory usage and computational time. This factor must therefore be considered in large-scale applications or Big Data contexts.

E. Practical Implications

In a biomedical context such as the one investigated, methodological choice depends primarily on the analytical objective:

- For a global descriptive analysis and rapid interpretation of principal axes (e.g., maternal profile versus neonatal profile), the Center method is appropriate.
- For analyses focused on measurement variability (clinical uncertainty, inter-visit fluctuations), the TOPS method provides a more faithful representation.

Thus, the results do not identify a universally superior method, but rather establish a decision framework depending on:

- The level of uncertainty,
- The average width of the intervals,
- The desired trade-off between interpretability and geometric fidelity.

In summary, Center-based PCA constitutes a robust, stable, and interpretable approach suitable for routine exploratory analyses. In contrast, Vertex-based PCA (TOPS), although computationally more demanding and more complex to interpret, offers a more comprehensive representation of the internal structure of interval-valued symbolic data. This complementarity suggests that methodological choice should be guided by the nature of the underlying uncertainty and the analytical objectives pursued.

VII. LIMITATIONS

A. Methodological Limitations

Despite the theoretical and empirical contributions of this study, several limitations should be acknowledged in order to properly situate the scope of the results.

B. Dimensional Inflation In The Tops Method

The main structural limitation of the Vertex-based method (TOPS) lies in the increase in data dimensionality. For a dataset containing ppp interval-valued symbolic variables, the TOPS transformation results in a $2p - dimensional\ representation$. This duplication of variables (lower and upper bounds) leads to:

- A quadratic increase in computational cost,
- Enlargement of the covariance matrix,
- Higher memory consumption.

From a theoretical standpoint, if classical PCA has complexity $O(np^2)$, the TOPS method reaches:

$$O(n(2p)^2) \; soit \; 4O(np^2).$$

In high-dimensional contexts, this inflation may become a limiting factor. It affects not only computation time but also numerical stability during spectral decomposition.

Therefore, although TOPS better preserves the internal structure of intervals, its large-scale applicability remains conditioned by available computational resources.

C. Difficulties In Factorial Interpretation

A second limitation concerns the interpretation of factorial axes.
In the Center method, each individual is represented by a single point, which facilitates:

- Reading of correlation circles,
- Identification of dominant variables,
- Substantive interpretation of principal components.

By contrast, in the TOPS method:

- Each individual is implicitly represented by a set of vertices,
- Variance is distributed across a larger number of components,
- Vectors in the correlation circle are shorter and more diffuse.

This dispersion makes interpretation more delicate, particularly when intervals are wide or heterogeneous.
Moreover, the extracted axes may simultaneously reflect:

- Inter-individual variability,
- Intra-individual variability.

Disentangling these two sources of variance requires additional analytical tools beyond the scope of the present study.

D. Euclidean Metric Assumption

In our implementation, the covariance matrix is based on Euclidean distance. However, for symbolic data, this assumption may be restrictive. Intervals possess a specific geometric structure that could be better captured using alternative metrics such as:

- Hausdorff distance,
- Wasserstein (optimal transport) distance,
- Probabilistic metrics adapted to complex objects.

The exclusive use of the Euclidean metric therefore constitutes a simplifying assumption that could be enriched in future work.

VIII. FUTURE PERSPECTIVES

The identified limitations open several promising research directions.

A. Kernel Symbolic Pca

A first perspective consists in extending the approach toward nonlinear symbolic PCA. Kernel PCA projects data into a potentially infinite-dimensional Hilbert space via a kernel function:

$$K(x_i, x_j) = \phi(x_i) \cdot \phi(x_j).$$

Adapting this framework to interval-valued symbolic data could:

- Capture nonlinear structures,
- Model complex relationships between bounds,
- Improve group separation in the factorial space.

Such an extension would be particularly relevant when intervals reflect dynamic or nonlinear phenomena (e.g., medical trajectories or fluctuating environmental measurements).

B. Wasserstein-Based Covariance Matrices

A second perspective concerns the construction of covariance matrices based on distances specifically adapted to intervals. The Wasserstein distance (optimal transport distance) measures the distance between distributions or intervals while accounting for internal structure.

Applied to intervals, it would allow:

- Simultaneous integration of position and width,
- Distinction between intervals with identical centers but different dispersions,
- Construction of a more representative similarity matrix.

A PCA based on a Wasserstein dissimilarity matrix would constitute a robust and geometrically coherent extension of symbolic PCA.

C. Monte Carlo Approximation Of Vertices For Scalability

To mitigate TOPS dimensional inflation, a possible strategy consists of stochastic vertex approximation. A Monte Carlo approach could:

- Reduce effective dimensionality,
- Lower computational cost,
- Preserve a faithful approximation of internal variability.

This would enable application of the TOPS method to high-dimensional datasets.

D. Extensions Toward Dynamic Symbolic Data

A natural extension concerns longitudinal symbolic data.
In many contexts (biomedical, climatological, industrial), intervals evolve over time. A dynamic symbolic PCA could integrate:

- Temporal variation of bounds,
- Correlation between successive intervals,
- Modeling of interval trajectories.

Such an approach would extend symbolic PCA to complex dynamic systems. In conclusion, although Center-based and Vertex-based symbolic PCA provide a solid methodological foundation for interval-valued data analysis, several theoretical and computational challenges remain. The proposed perspectives—kernel PCA, Wasserstein-based metrics, Monte Carlo approximations, and dynamic extensions—offer a fertile framework for the future development of more robust, scalable, and geometrically faithful symbolic dimensionality reduction methods.

X. CONCLUSION

In this work, we addressed the problem of dimensionality reduction for interval-valued symbolic data, which are increasingly encountered in biomedical, environmental, industrial, and socio-economic domains. These data, characterized by internal variability and structural uncertainty, require rigorous adaptation of classical factorial methods, particularly Principal Component Analysis (PCA).
The comparative analysis between Center-based PCA and Vertex-based PCA (TOPS) reveals a fundamental trade-off between dimensional condensation and preservation of internal variability. The Center method, based on midpoint transformation, offers several advantages:

- Strong variance concentration on the first axes,
- Clear interpretability of principal components,
- Stable variable contributions,
- Moderate computational cost.

It therefore constitutes a robust and operational solution when intervals are relatively homogeneous or when the primary objective is exploratory synthesis. By contrast, the TOPS method, through explicit exploitation of lower and upper bounds, more faithfully preserves the internal geometry of symbolic objects. It becomes particularly relevant when:

- Interval widths are substantial,
- Uncertainty is intrinsic to the analysis,
- Intra-object dispersion must be retained.

However, this descriptive richness results in:

- More dispersed variance distribution,
- More complex factorial interpretation,
- Increased computational cost.

Thus, neither method can be considered universally superior. Their relevance depends on application context, uncertainty level, and analytical objectives. Beyond empirical comparison, this study contributes a structured methodological framework for evaluating symbolic PCA extensions and opens promising research directions toward nonlinear approaches, adapted metrics, and complexity-reduction strategies. Ultimately, symbolic PCA for interval-valued data remains an evolving research field. The development of theoretically sound, computationally efficient, and interpretable methods constitutes a central challenge in addressing the growing complexity of contemporary data.

REFERENCES

[1] L. Billard, Sample Covariance Functions for Complex Quantitative Data: Conference Yokohama Japan 2008
[2] L. Carlo and P. Fabrizia, Principal component analysis of interval data: a symbolic data analysis approach .
[3] L. Billard et E. Diday, Symbolic Data Analysis: Conceptual Statistics and Data Mining .
[4] H. Vilela, S. Dias and P. Brito, Extracting information from interval data using symbolic principal component analysis: Communications in Statistics Simulation and Computation .
[5] F. Wang, J. Chen and X. Gong, An Improved Interval-type Symbolic Data Principal Component Analysis .
[6] P. Cazes, A. Chouakria, E. Diday et Y. Schektman. Extension de l'analyse en composantes principales à des données de type intervalle .
[7] N. Carlo, L. Billard and F. Palumbo, Principal Component Analysis of Interval Data: a Symbolic Data Analysis Approach .
[8] E. Diday, Introduction à l'approche symbolique en analyse des données : Journées Symbolique-Numerique .
[9] E. Diday, Une introduction à l'analyse des données symboliques, SFC .
[10] G. Meccariello, Analisi in componenti principali per dati ad intervallo.
[11] I. Jolliffe and J. Cadima, Principal component analysis: a review and recent developments.
[12] Golub and V. Loan, Principal component analysis: a review and recent developments.
[13] E. Diday and F. Esposito, An introduction to Symbolic Data Analysis and the Sodas Software IDA.
[14] E. Diday, Symbolic Data Analysis: Past, Present and Future
[15] F. Noirhomme and P. Brito, Far beyond the classical data models: Symbolic data analysis.
[16] F. Hausdorff, Grundzüge der Mengenlehre. Leipzig, Veit & Comp. Classical reference for the Hausdorff distance
[17] H-H Bock and E. Diday, *Analysis of Symbolic Data: Exploratory Methods for Extracting Statistical Information from Complex Data*
[18] A. Irpino and R. Verde, Basic statistics for distributional symbolic variables: A new metric-based approach. *Advances in Data Analysis and Classification*
[19] E. Diday and J-C Simon, Clustering analysis in *digital Pattern Recognition*